Desviación Estándar

Cuando tenemos como en la gráfica como están distribuidos los datos, donde el eje x; es el precio de los autos. Y el precio **promedio** de los autos es el punto rojo. Decimos que una buena medida de dispersión que no es el rango, ni el rango intercuartil, viene dada por **calcular la variación** de cada punto (los puntos verdes) y ver cual es la distancia que tiene ese punto respecto del valor promedio, entonces recuerda el valor **promedio** está en rojo. Entonces algunas veces esa distancia es negativa, positiva, para evitar este tipo de situación y poder sumar las contribuciones que hacen las distancias con respecto del valor promedio, lo que se suele hacer es; tomar cada punto x_i compararlo con μ el valor promedio y elevarlo al cuadrado.

$$(x_i - \mu)^2$$

Cuando se hace esto evito el problema de que existan distancias positivas y negativas, dicho esto. **La desviación estándar** viene dada por un concepto llamado **varianza**

$$varianza = \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

$$desviación\ estándar = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

Hay algo importante a mencionar, cuando nosotros tenemos un conjunto de datos, ese conjunto de datos normalmente suele ser una $muestra$ (dibujo en azul) de lo que realmente es toda la población del problema que estamos analizando, entonces en estadística hablamos que existe la *desviación estándar de la población* y la *desviación estándar de la muestra*.

Cuando vemos que se calculan respecto de la muestra podemos ver que se calcula de diferente forma $N = n - 1$ lo siguiente:

$$\sigma_{sample} = \sigma_{muestra} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{N}(x_i - \mu)^2}$$

La explicación es que hay razones matemáticas de que $n - 1$ sea un factor de corrección cuando se esta trabajando con una muestra de datos y no con toda la población. En este curso no se profundizará en esto, pero hay que tener en cuenta este aspecto.

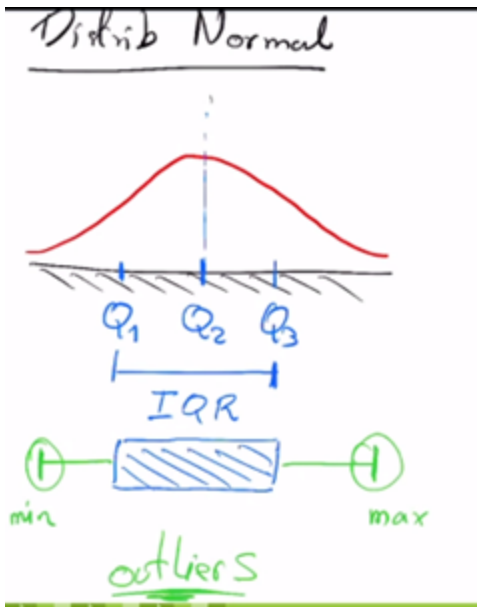
Recordando

$$\{x_1, x_2, x_3, .., x_n\}$$

$$\bar{x} = \mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Distribución normal o Gaussiana

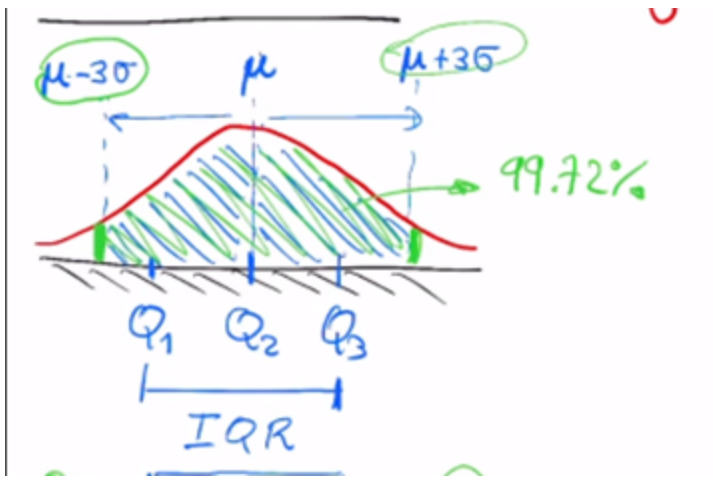
Vamos pensar en la distribución de datos como forma de campana en lugar de las barras o histograma. Cuando tenemos una distribución normal, justo el valor de la mitad es la ***mediana*** y que justo también coincide con la ***mediana(Q2)*** y alrededor de el existe el **Q1** y **Q3** y tenemos que a eso le llamamos **IQR** o **Rango Intercuartil** y de ahí asociamos el **Box plot** que contiene el **Q1,Q2 y Q3** y normalmente se le dibujan unos brazitos que dan una idea de que tan lejos está localizada o esparcida la distribución de datos de los valores **(min,max)**. Pero normalmente estos extremos se suelen ajustar para que nos permitan excluir **datos anómalos** u **Outliers**.



A todo esto ¿donde cobra sentido la desviación estándar σ

Lo que sucede es que si nosotros de nuevo estamos asumiendo una distribución normal como la de la gráfica, que es simétrica, es decir está perfectamente balanceada de la media. Lo que sucede es que el σ también mide desviaciones, para el caso de la distribución normal, cuando yo considero alrededor de la **(media y mediana)** que coinciden para el caso de una distribución normal **3 distancias adelante y atrás de la media** (μ), la teoría de la distribución normal nos dice que

> Cuando yo considero todos los datos entre $\mu - 3\sigma$ y $\mu + 3\sigma$ ahí estoy contemplando básicamente la totalidad de los datos ≈ 99.72 de todos los datos de esa distribución.

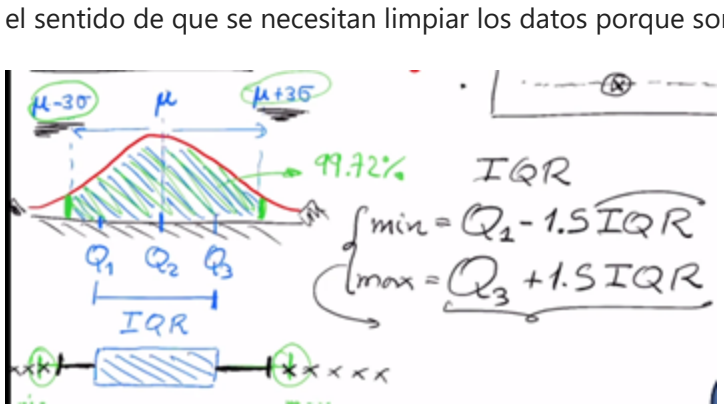


Esto quiere decir que si yo tengo un conjunto de datos y a los costados tengo datos mas lejanos, nos puede decir que son **datos anómalos u outliers** y que están lo suficientemente lejos $\approx \mu \pm 3\sigma$.

Entonces hay un método para detectar estos datos anómalos y es basado en el **IQR** o **Rango Intercuartilico**

$$rango\ de\ datos\ normales = \begin{cases} min = Q_1 - 1.5IQR \\ max = Q_1 + 1.5IQR \end{cases}$$

Entonces el limite del **Box plot se ajusta** para eliminar los **datos anómalos(Taches negros)**. Entonces todo lo que este por fuera de estas barras lo identifico como datos anómalos, y que normalmente en procesos estadísticos son datos que son descartados, en el sentido de que se necesitan limpiar los datos porque son datos que no van acorde al patrón con el que se quiere identificar.

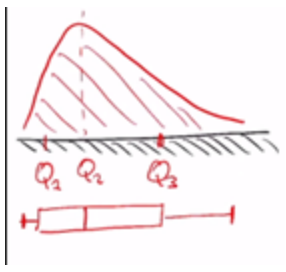


A LO ANTERIOR SE LE LLAMA MÉTODO DE DETECCIÓN DE OUTLIERS CON RANGO INTERCUARTIL

El valor de 1.5 en el proceso anterior coincide con $\mu - 3\sigma$, entonces coincide que en ambos estamos considerando el 99.72 de datos normales, coinciden y son muy cercanos a que estamos tomando 3 desviaciones estándar respecto del valor medio, no es exacto pero es aproximado. Así podemos decir que es una justificación para tomarlo de referencia.

¿Pero que pasa en el caso de la distribución no geométrica?

Digamos que la media y los cuartiles no estén simétricamente separados, entonces tendría una representación diferente en **box plot**. En este caso no aplicaría el criterio:



$$rango\ de\ datos\ normales = \begin{cases} min = Q_1 - 1.5IQR \\ max = Q_1 + 1.5IQR \end{cases}$$

Aquí lo que se hace es tratar de hallar los puntos en los cuales los datos se consideran anómalos.

Lo que suele hacer realmente:

$$Hallando \quad \begin{cases} Q_1 - 1.5\ f(IQR) \\ Q_3 + 1.5\ g(IQR) \end{cases}$$

Es decir se toman en cuenta funciones que dependan de IQR. Y la idea de colocar estas funciones es tomar en consideración ese sesgo de la distribución asimétrica y que entonces se pueda calcular la longitud de los brazos. De forma que se ajusta perfectamente a la distribución real

Extras:

Aporte de Bryan

Un error común es comparar la desviación estándar de dos variables que tienen escalas diferentes, en este ejemplo usaremos el edad y altura edad cumplida media = 24, desviación estándar = 2.5 altura en centimentros media = 175, desviación estándar = 8 Como vemos la desviación estándar de altura es mayor dado a su escala.

Lo correcto sería usar el coeficiente de variabilidad que es simplemente dividir la desviación estandar entre la media.

edad 2.5 / 24 = 0.1042.

altura 8 / 175 = 0.0457

Ahora ambas variables la podemos compara y concluimos que edad tiene una mayor variabilidad

> A este proceso se le llama estandarización, o bueno, así lo conozco, y sirve para transformar los datos con el objetivo de tener una media de 0 y una desviación estándar en función a esta (<1).(Jose Eduardo Victorio Gonzales)

Miguel Angel Velazquez Romero

Desviación estándar La desviación estándar es la medida de dispersión más común, que indica qué tan dispersos están los datos con respecto a la media. Mientras mayor sea la desviación estándar, mayor será la dispersión de los datos. El símbolo σ (sigma) se utiliza frecuentemente para representar la desviación estándar de una población, mientras que s se utiliza para representar la desviación estándar de una muestra. La desviación estándar se puede utilizar para establecer un valor de referencia para estimar la variación general de un proceso.

.

Solo para aclarar, ya que el termino de varianza se abordo muy rápidamente: .

Varianza: es una medida de dispersión que representa la variabilidad de una serie de datos respecto a su media. Formalmente se calcula como la suma de los residuos al cuadrado divididos entre el total de observaciones. Su fórmula es la siguiente: X → Variable sobre la que se pretenden calcular la varianza xi → Observación número i de la variable X. i puede tomará valores entre 1 y n. N → Número de observaciones. $\bar{x}$ → Es la media de la variable X. La diferencia entre la desviación estándar o típica y la varianza, es que la la desviación típica es la raíz cuadrada de la varianza . Y no nos podemos olvidar de otra medida de dispersión muy importante, que es el coeficiente de variación:

Su cálculo se obtiene de dividir la desviación típica entre el valor absoluto de la media del conjunto y por lo general se expresa en porcentaje para su mejor comprensión.

X → Variable sobre la que se pretenden calcular la varianza σx → Desviación típica de la variable X. | $\bar{x}$ | → Es la media de la variable X en valor absoluto con $\bar{x} \neq 0$ El coeficiente de variación de utiliza para comparar la dispersión (variación) de conjuntos de datos de medidas diferentes o con medias aritméticas diferentes.